

Weekly Report

December 20, 2014

EM algorithm

想看懂EM算法很久了，这周算是有了一个大概的了解。我先后看了《统计学习基础》、《The Elements of Statistical Learning》和Andrew Ng 斯坦福课程及材料。其中《The Elements of Statistical Learning》讲得最简短，《统计学习基础》提到了一些算法的重点，却没有说明是为什么，Andrew Ng讲解最为详细。同时算法的导出和算法的证明，《统计学习基础》、《The Elements of Statistical Learning》分别使用了公式(1)和(2)，Andrew Ng则都使用了公式(1)，相对于初学者容易理解。这次周报也是写下我学习中的体会，便于以后翻阅。

$$P(Y|\theta) = \sum_Z P(Y, Z|\theta) = \sum_Z P(Y|Z, \theta)P(Z|\theta) \quad (1)$$

$$P(Y|\theta) = \frac{P(Y, Z|\theta)}{P(Z|Y, \theta)} \quad (2)$$

有时训练数据只有输入没有对应的输出 $\{(x_1, \cdot), (x_2, \cdot), \dots, (x_N, \cdot)\}$ ，从这样的数据学习模型称为非监督学习问题，EM算法可以用于生成模型的非监督学习。假设有N组独立的观测数据 x_1, x_2, \dots, x_N ，隐变量数据Z，想要满足模型 $P(X, Z|\theta)$ 。我们面对一个含有隐变量的概率模型，目标是极大化观测数据X关于参数 θ 的对数似然函数，即最大化

$$\begin{aligned} l(\theta) &= \sum_{i=1}^m \log p(x_i; \theta) \\ &= \sum_{i=1}^m \log \sum_{z_j} p(x_i, z_j; \theta) \end{aligned} \quad (3)$$

这一极大化的主要困难是 \log 中的存在和(或积分)，比如 $\log(e^x + e^y)$ 求极大值比较困难，我们希望目标函数是类似于 $\log(e^x) + \log(e^y)$ 的形式。那么是否可以交换 \log 和 \sum (也可以是求期望 E)?下面需要一个不等式，描述了凹函数凸函数的性质。

Jensen inequality

f 是一个在实数域上的函数，如果满足 $f''(x) \geq 0, \forall x \in \mathbb{R}$ ，则称 f 是凸函数；如果 $f''(x) > 0, \forall x \in \mathbb{R}$ ，则称 f 是严格凸函数。

定理：如果 f 是一个凸函数， X 是随机变量，那么 $E[f(x)] \geq f(E[x])$

推论：如果 f 是一个严格凸函数，当且仅当 $x = E(x)$ 时， $E[f(x)] = f(E[x])$ 。

比如， x 是一个常数。

推论：如果 f 是一个凹函数，Jensen inequality也成立，符号要变一下方向 ($E[f(x)] \leq f(E[x])$)。

*函数图可以查看Andrew Ng的课程材料

EM algorithm 导出

对数似然函数中 \log 是一个严格凹函数，根据Jensen inequality $E[f(x)] \geq f(E[x])$ ，我们希望找到 $l(\theta)$ 的一个下界，并且是紧的（尽可能接近 $l(\theta)$ ，即使得等号成立）。

$$\begin{aligned} l(\theta) &= \sum_{i=1}^m \log \sum_{z_j} p(x_i, z_j; \theta) \\ &= \sum_{i=1}^m \log \sum_{z_j} Q_i(z_j) \frac{p(x_i, z_j; \theta)}{Q_i(z_j)} \end{aligned} \quad (4)$$

把 \sum_{z_j} 看做求期望，把 $Q_i(z_j)$ 看做是概率，自然的，这里要求 $\sum_{z_j} Q_i(z_j) = 1, Q_i(z_j) > 0$ 。

$$\begin{aligned}
l(\theta) &= \sum_{i=1}^m \log \sum_{z_j} Q_i(z_j) \frac{p(x_i, z_j; \theta)}{Q_i(z_j)} \\
&= \sum_{i=1}^m \log E_{z_j} \left[\frac{p(x_i, z_j; \theta)}{Q_i(z_j)} \right] \\
&\geq \sum_{i=1}^m E_{z_j} \left[\log \frac{p(x_i, z_j; \theta)}{Q_i(z_j)} \right] \\
&= \sum_{i=1}^m \sum_{z_j} Q_i(z_j) \log \frac{p(x_i, z_j; \theta)}{Q_i(z_j)} \\
&= J(Q, \theta)
\end{aligned} \tag{5}$$

要求 $l(\theta)$ 的最大，可以通过求 $J(Q, \theta)$ 的最大，求 $J(Q, \theta)$ 的最大可以看做坐标上升算法，固定 θ 求 Q ，再固定 Q 求 θ 。固定 θ 求 Q 的过程其实就是求 $l(\theta)$ 最大下界的过程。刚才 Jensen inequality 讨论中说道，如果 $\frac{p(x_i, z_j; \theta)}{Q_i(z_j)}$ 是一个常数，等号就成立，下界就是紧的。

$$\begin{aligned}
\frac{p(x_i, z_j; \theta)}{Q_i(z_j)} &= c \\
Q_i(z_j) &\propto p(x_i, z_j; \theta)
\end{aligned} \tag{6}$$

那么考虑到 $\sum_{z_j} Q_i(z_j) = 1$ （因为这是一个概率分布）， $Q_i(z_j)$ 就可以这样取值：

$$\begin{aligned}
Q_i(z_j) &= \frac{p(x_i, z_j; \theta)}{\sum_{z_j} p(x_i, z_j; \theta)} \\
&= \frac{p(x_i, z_j; \theta)}{p(x_i; \theta)} \\
&= p(z_j | x_i; \theta)
\end{aligned} \tag{7}$$

这样选择的 Q_i 使得 $l(\theta)$ 的下界最大($Q = \arg \max_Q J(Q, \theta)$), 此时 $l(\theta) = J(Q, \theta)$ 。确定 Q_i 就是EM算法的E步-求期望。在M步中，我们就要关于 θ 最大化下界

$$\theta = \arg \max_{\theta} J(Q, \theta) \tag{8}$$

EM算法是一个迭代过程，如上可以不断通过 θ 求得新的 θ' ，直到收敛。

EM算法描述如下：

迭代至收敛{

(E)对每一个i,

$$Q_i(z_j) := p(z_j|x_i; \theta)$$

(M)更新 θ ,

$$\theta := \arg \max_{\theta} \sum_{i=1}^m \sum_{z_j} Q_i(z_j) \log \frac{p(x_i, z_j; \theta)}{Q_i(z_j)}$$

}

EM algorithm 证明

这里证明一部分算法为什么会收敛。假设 θ^t 和 θ^{t+1} 是一次迭代前后 θ 的值，下面证明 $l(\theta^t) \leq l(\theta^{t+1})$ 。

算法从 θ^t 出发，我们会使得 $Q_i^t(z_j) := p(z_j|x_i; \theta^t)$ ，那么，

$$\begin{aligned} l(\theta^t) &= J(Q^t, \theta^t) \\ &= \sum_{i=1}^m \sum_{z_j} Q_i^t(z_j) \log \frac{p(x_i, z_j; \theta^t)}{Q_i^t(z_j)} \end{aligned} \quad (9)$$

θ^{t+1} 是最大化(9)的右边得到的，所以，(*《统计学习基础》P160在图上做了四个点的位置大小关系的示意图)

$$l(\theta^{t+1}) \geq \sum_{i=1}^m \sum_{z_j} Q_i^t(z_j) \log \frac{p(x_i, z_j; \theta^{t+1})}{Q_i^t(z_j)} \quad (10)$$

$$\geq \sum_{i=1}^m \sum_{z_j} Q_i^t(z_j) \log \frac{p(x_i, z_j; \theta^t)}{Q_i^t(z_j)} \quad (11)$$

$$= l(\theta^t) \quad (12)$$

第一个不等号(10)是因为，

$$l(\theta) \geq \sum_{i=1}^m \sum_{z_j} Q_i(z_j) \log \frac{p(x_i, z_j; \theta)}{Q_i(z_j)}$$

不等号(11)是因为M步，

$$\theta^{t+1} := \arg \max_{\theta} \sum_{i=1}^m \sum_{z_j} Q_i^t(z_j) \log \frac{p(x_i, z_j; \theta)}{Q_i^t(z_j)}$$

等号(12)是因为E步。

以上证明了 $l(\theta)$ 是递增的。如果 $l(\theta)$ 有上界，就会收敛到某一值 $l(\theta^*)$ ，如果 $Q, l(\theta)$ 满足一定条件也会收敛到稳定点，《统计学习基础中提到》。

Plan for next week

这周在EM算法上花费了许多时间，下周会把重点调整到轨迹数据可视化论文的阅读上。